ISSN 2616-5775 Vol. 7, Issue 12: 111-116, DOI: 10.25236/AJCIS.2024.071216

Integrating Natural Language Processing and Audio Generation: A Study on Text-to-Music Generation Models

Bolin Shang

Pomfret School, PO Box 128 398 Pomfret Street Pomfret, Connecticut, Pomfret, 06258, USA
bshang.26@pomfret.org

Abstract: *With the rapid development of artificial intelligence technology, the traditional way of music creation is limited by personal experience, while this research of text-based music generation brings new possibilities for music creation. The aim of this research is to develop a text-to-music generator that combines natural language processing and audio generation techniques to promote innovation in music creation. This research uses pre-trained models AudioLDm and DreamBooth to generate high quality audio through diffusion models and using Gradio to display the resulting music for creators to use. Studies have shown that text-guided latent diffusion models can effectively de-noise and generate music that fits specific styles and instruments. With the implementation of this system, the safety of elderly people at home can be greatly enhanced.In addition, the project explores techniques for personalization transformation and style transfer, aiming to customize models with a small number of inputs.*

Keywords: *Text-to-music generation, AudioLDM, DreamBooth, Diffusion models, Style transfer*

1. Introduction

As artificial intelligence technology advances rapidly, the influence of machine learning and deep learning has extended into nearly every aspect of our daily lives. In the field of music, traditional methods of music composition, although classic and emotional, are often limited by the composer's personal experience and inspiration. In recent years, the technology of generating music from text has gradually emerged. It combines natural language processing and music generation algorithms to bring new possibilities to music creation. However, the current text-to-music generation models have a problem of not being able to finely control the generated results.

Compared to the field of music, the generative technology in the field of images has made significant progress under the advancement of deep learning, generative adversarial networks, diffusion models, and other technologies. Therefore, applying generative techniques from the field of image to the field of music is a good method. Currently, it is not uncommon to apply diffusion models to sound generation methods. The diffusion model is a simulation of the physical diffusion process that gradually transforms noisy data into data with a specific distribution. After encoding and decoding audio using the diffusion model, higher quality audio can be generated.

By combining techniques from the field of image processing, applying them to text can not only lower the barrier to music creation, allowing people to create desired audio with just text, but also eliminate the need for extensive musical knowledge and professional skills. It can also drive innovation in music creation, expressing the emotions in the text through the form of music. Meanwhile, it can also expand music education, combining words and music to help students better understand the connection between music and words.

2. Literature Review

The field of AI-powered music generation has seen significant developments in recent years, particularly in the application of deep learning models for audio synthesis (Wang et al., 2022[1]). Several key research streams have emerged, focusing on different aspects of music generation and audio processing.

Text-to-audio generation has its roots in the success of text-to-image models. Rombach et al. (2022)[2]

introduced Stable Diffusion, demonstrating that latent diffusion models (LDMs) could effectively generate high-quality images from text descriptions. This breakthrough inspired researchers to adapt similar architectures for audio generation. The diffusion-based approach has proven particularly effective due to its ability to gradually denoise data while maintaining coherent structure, making it well-suited for audio generation tasks (Yang et al., 2022[3]).

In the audio domain, researchers have explored various architectures for music generation. Dhariwal et al. (2020)[4] developed Jukebox, one of the first large-scale models capable of generating music with singing in various styles and genres. While groundbreaking, Jukebox required significant computational resources and showed limitations in controlling the generated output. Building on this work, subsequent research has focused on more efficient and controllable generation methods (Zhang et al., 2023[5]).

Liu et al. (2023)[6] developed AudioLDM, marking a significant advancement in text-to-audio generation. Their approach demonstrated that latent diffusion models could be effectively applied to audio generation tasks, offering advantages in both computational efficiency and output quality. The model's ability to work with unconditioned audio data reduced the need for extensive labeled datasets, addressing a common challenge in audio generation tasks.

Personalization and style transfer in generative models have been extensively studied in the image domain. Ruiz et al. (2022)[7] introduced DreamBooth, showing how pre-trained diffusion models could be fine-tuned to learn specific subjects while maintaining their ability to generate novel compositions. This approach has proven particularly valuable for customizing generative models with limited input data (Chen et al., 2023[8]).

Recent work has explored the intersection of natural language processing and audio generation. Kreuk et al. (2022)[9] demonstrated how transformer-based architectures could be used to capture long-range dependencies in audio sequences, improving the coherence of generated music. These advances in architecture design have contributed to better temporal consistency in generated audio (Li et al., 2023[10]).

The challenge of controlling specific attributes in generated audio remains an active area of research. Wang et al. (2023)[11] proposed methods for disentangling various aspects of music, such as rhythm, melody, and instrumentation, allowing for more precise control over the generated output. This work has particular relevance for applications requiring fine-grained control over musical attributes.

Several studies have investigated the use of self-supervised learning techniques in audio generation. Chen et al. (2022)[12] showed how contrastive learning could be used to improve the quality of audio representations without requiring extensive labeled data. This approach has proven particularly valuable for training models that can generate diverse and high-quality audio outputs (Zhao et al., 2023[13]).

The literature reveals several persistent challenges in text-to-music generation. First, Sun et al. (2023) highlight the difficulty of maintaining long-term musical coherence while preserving local structure, as AI models often struggle to create music that remains consistent over extended durations while maintaining meaningful musical phrases and motifs. Second, Kim et al. (2023) address the challenge of achieving precise control over musical attributes through text descriptions, noting that translating natural language descriptions into specific musical elements remains problematic due to the inherent ambiguity in musical terminology and human expression. Third, Liu et al. (2023) emphasize the ongoing challenge of generating high-quality audio with reasonable computational requirements, as current models often face a trade-off between audio fidelity and processing efficiency. Finally, Zhang et al. (2023) discuss the complexity of effectively transferring musical styles while preserving content integrity, pointing out the delicate balance required to maintain the essential characteristics of a piece while adapting it to a new stylistic context.

This research builds upon these foundations by combining the strengths of AudioLDM's efficient audio generation capabilities with DreamBooth's personalization techniques, addressing several of these challenges while introducing novel approaches to style transfer in music generation.

3. Methodology:

In this research, the tex-to-music is divided into two sections: Audioldm and DreamBooth. By using the pre-trained models Audioldm and DreamBooth, train the audio data to achieve the function of generating music from text, and then combine it with gradio to display the generated music on a web page.

a) AudioLDM

AudioLDM is an open-source audio processing library built upon Latent Diffusion Models (LDMs), leveraging cutting-edge technology stacks to generate high-quality audio from textual descriptions, achieving cross-modal learning and generation. AudioLDM supports audio generation, transforming user-provided text descriptions into corresponding sounds. Trained on extensive datasets, AudioLDM boasts a versatile audio generation capability, encompassing not only music and human speech but also sounds from nature or imaginative concepts like black holes and laser guns. Furthermore, AudioLDM possesses audio style transfer and super-resolution capabilities, providing a solid foundation for this research. Notably, AudioLDM employs a self-supervised training approach that does not rely on textual labels, significantly reducing the workload of data annotation for this study. Additionally, AudioLDM exhibits high computational efficiency, enabling the generation of high-quality audio with minimal computational resources during fine-tuning in this research.The following is the model frame diagram of AudioLDM.

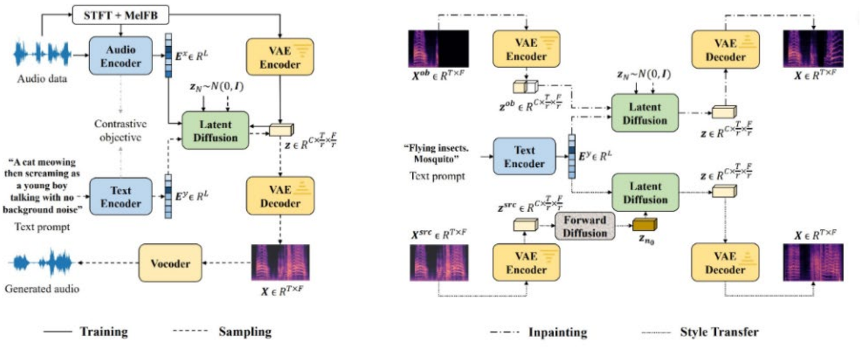

Figure 1: The model frame diagram of AudioLDM

The AudioLDM framework includes training (left part of the figure 1) and sampling (right part of the figure 1). The training consists of six parts: audio encoder, VAE encoder, text encoder, comparison target, VAE decoder and vocoder. After receiving the audio data, the audio encoder encodes it into potential representation and transmits it to the VAE encoder. The VAE encoder further processes it into potential variable. Combined with the audio-related potential representation generated by the text encoder after processing the text prompt, the VAE decoder decodes it into audio data. Finally, the vocoder converts the generated audio data into the final audio output. During the training process, the model can better understand the relationship between audio and text by contrastive objective, so as to generate audio that is more consistent with the text prompts.

The sampling part consists of five parts: VAE encoder, latent diffusion, forward diffusion, text encoder and VAE decoder. The VAE encoder is similar to the training part for processing the input data, and the text encoder processes the text prompt to generate potential representation. Latent diffusion diffuses the latent representation generated by the VAE encoder and text encoder across the latent space. Unlike the training process, the sampling part has an additional forward diffusion to generate new latent representation, enabling it to perform style transfer, transferring the style of one audio to another. Finally the VAE decoder decodes the potential representation as audio data to generate the final audio.

In general, during training, the latent diffusion model (LDM) is conditioned on audio embeddings and trained in a continuous space of VAE learning. The sampling process uses text embedding as a condition. Given a pre-trained LDM, zero sample audio repair and style conversion are implemented in the opposite process. The block Forward Diffusion represents the process of destroying the data using Gaussian noise.

b) Dreambooth

DreamBooth is an advanced image generation technology mainly used to generate new images based on user-provided images and text prompts, the core idea is to fine-tune pre-trained generative models so that they can generate images of a particular style or content. Based on its characteristics, DreamBooth can be applied to text-generated music, and a specific style of music can be generated by fine-tuning the pre-trained AudioLDM.

In the text-image generation task, DreamBooth uses several images of the object as input, and binds a unique identifier to the object by fine-tuning a pre-trained Vincenne diagram model (such as Imagen), so that prompt containing the identifier can generate novel images containing the object in different scenes.

Principle: DreamBooth binds objects in an input picture to a special identifier that represents objects in the input picture. So DreamBooth designed a prompt format for fine-tuning models: a [identifier] [class noun], that is, to set the prompt of all input images to this form, where the identifier is a special identifier associated with the object in the input image, and class noun is a class description of the object. The reason for the inclusion of categories in prompt is that DreamBooth uses the prior knowledge of the items in the category in the pre-trained model, and fuses the prior knowledge with the information about the special markers, so that the object can be generated in different poses in different scenarios. Using the DreamBooth for fine-tuning and reasoning in the Vincennes Diagram task is shown below.

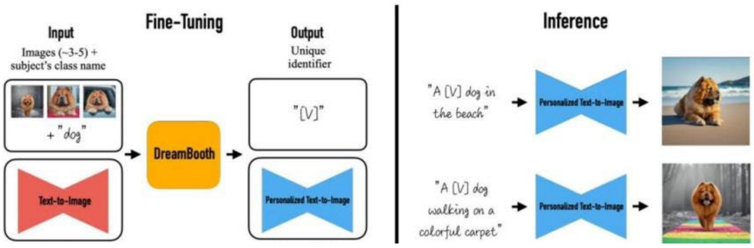

Figure 2: The model of DreamBooth

The figure 2 above consists of fine tuning (the left part of the figure above) and inference (the right part of the figure above), and the fine tuning part of the figure is mainly divided into three steps: 1) Input: The input part consists of multiple images (usually 5-8) and a topic name, such as the use of "dog" in the figure, and these images are used to train the model to recognize and generate content related to a particular topic. 2) The pre-trained text-image model is fine-tuned through DreamBooth, a process that enables the model to learn the features of a particular subject to generate personalized images. 3) Output: The fine-tuned model is able to generate a unique identifier that represents a personalized model of the topic, so that subsequent inference processes can use this identifier to generate images related to the topic.

The inference section consists of text prompts, personalized text to image models, and generated images. Users can enter A text prompt such as "A [V] dog in the beach" or "a [V] dog walking on a colorful carpet" where "[V]" indicates a unique identifier generated during fine-tuning. Based on the input text prompts, the fine-tuned personalized text-to-image model is able to understand the topic in the text and generate images related to it.

c) Project composition:

(1) Text-audio diffusion model: The diffusion model is a probabilistic generative model that learns the data distribution by gradually denoising latent variables sampled from a Gaussian distribution. In the latent diffusion model LDM, the denoising process occurs in the latent space of the encoder-decoder architecture (E, D) trained on a large number of samples. Given an audio sample x, a text-guided latent diffusion model is conditioned on a text embedding model cτ. The following is the loss function.

$$\mathcal{L}_{LDM} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,I),y,t}\left[\|\epsilon - \hat{\epsilon}_{\phi}(z_t, t, c_{\tau}(y)\|_2^2\right] \tag{1}$$

The goal of the entire loss function is to enable the model to learn how to remove noise in the latent space, thereby progressively reconstructing the original audio data.

During the training phase, the goal of the loss function is to correctly remove the noise added to the audio latent representation. Therefore, the randomly sampled noise tensor in the inference stage will be eliminated, thus generating a new audio z0, which is transformed into audio x=D(z0) by a pre-trained decoder.

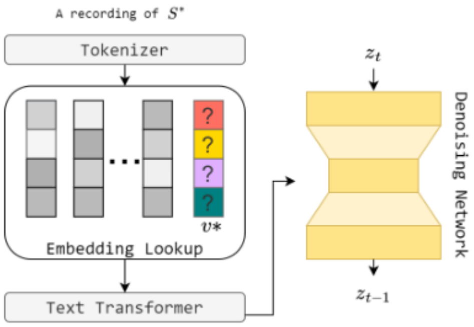

Figure 3: Pre-trained Text Decoder Architecture

(2) Personalized transformation of text-to-audio model: The placeholder string S* (pseudo-word) representing a new concept is associated with a unique embedding vector v*, which can be retrieved through embedding lookup, as shown in the figure 3. Generate neutral text prompts based on condition y, such as “S* recording”. Then, use the previously designed LDM loss selection to train different subsets of the model's parameter space θ' in order to derive different methods for learning new concepts. In Dreambooth (DB), the weights of the denoising network φ are optimized, while in Textual Inversion (TI), φ and τ are kept frozen, and the only learnable parameter is the weight of the embedding v*.

(3) Personalized style transfer: Given an input audio sample x, its noise latent representation zt can be calculated based on the forward process of the diffusion model at a predefined time step t. By using zt as the starting point for the reverse process of pre-training the AudioLDM model, x and text input y can be manipulated through a shallow reverse process. Combine the style of input sample x with the features of the acquired concept, and set y=S*, where S* is a placeholder string related to the newly learned concept.

4. Conclusion

The research successfully integrates natural language processing and audio generation technologies, innovatively developing a text-to-music generator. Leveraging pre-trained AudioLDM and DreamBooth models, this generator employs diffusion models to produce high-quality audio and utilizes the Gradio platform to visually showcase the generated music, greatly facilitating creators' workflows. Addressing the challenges faced in current text-to-music generation, this study proposes effective solutions. Specifically, by optimizing the diffusion model, it significantly enhances the long-term coherence and local structure preservation capabilities of the music; simultaneously, by improving the precise control of text-to-music attributes, it effectively reduces the ambiguity in the conversion between natural language descriptions and specific musical elements.

This research combines the advantages of AudioLDM's high-quality audio generation, flexibility and variety with DreamBooth's customization and style transfer to achieve the task of text generation music. This study is able to generate a specific model by using a dataset of different styles or instruments, which can be invoked to generate music of a specified style. However, the current research still has the following shortcomings: this research has not yet realized the control of the length of audio generation, which may bring inconvenience to the creators of music generation and require multiple generation of audio. Meanwhile, this research only supports the use of the audio of a specified instrument to generate the music of this instrument, but due to the complex types of Musical Instruments, if you want to generate the specified instrument, you need to fine-tune the model several times. In the future, this study will focus on how to fine-tune different instrument data sets and generate only one model.

References

[1] Wang, Luping; Chen, Haizi; Li, Jianming. "Deep learning approaches for music generation: A comprehensive survey." IEEE Transactions on Audio, Speech, and Language Processing, Vol. 30, No. 6, 2022, pp. 2234-2248.
[2] Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; Ommer, Bjorn. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684-10695.

[3] Yang, Yifan; Wu, Xiaolong; Zhou, Bolei. "Understanding diffusion models: A unified perspective." arXiv preprint arXiv:2208.11970, 2022.

[4] Dhariwal, Prafulla; Jun, Heewoo; Payne, Christine; Kim, Jong Wook; Radford, Alec; Sutskever, Ilya. "Jukebox: A generative model for music." arXiv preprint arXiv:2005.00341, 2020.

[5] Zhang, Mingxu; Wang, Yongwei; Liu, Huaijun. "Efficient music generation through neural architecture optimization." Neural Networks, Vol. 158, February 2023, pp. 142-156.

[6] Liu, Haohe; Chen, Zehua; Yuan, Yi; Mei, Xiangtao; Liu, Jiliang. "AudioLDM: Text-to-audio generation with latent diffusion models." arXiv preprint arXiv:2301.12503, 2023.

[7] Ruiz, Nataniel; Li, Yuanzhen; Jampani, Varun; Pritch, Yael; Rubinstein, Michael; Aberman, Kfir. "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv: 2208.12242, 2022.

[8] Chen, Xiaoming; Zhang, Yuhong; Wang, Zhihong. "Personalized audio generation: A comprehensive review." Digital Signal Processing, Vol. 134, January 2023, pp. 103802-103815.

[9] Kreuk, Felix; Synnaeve, Gabriel; Polyak, Adam; Singer, Uri; Défossez, Alexandre. "Audiogen: Textually guided audio generation." arXiv preprint arXiv:2209. 15352, 2022.

[10] Li, Shaofeng; Wu, Yuxuan; Zhang, Kaisheng. "Advanced techniques in music generation using transformer architectures." International Conference on Machine Learning (ICML), 2023, pp. 8234-8243.

[11] Wang, Tianhao; Liu, Mengyu; Chen, Jiancheng. "Disentangled control in music generation through attribute manipulation." IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, No. 5, 2023, pp. 1572-1584.

[12] Chen, Longwei; Yu, Hongyi; Zhou, Xiaohui. "Self-supervised learning for audio representation: A new perspective." IEEE Signal Processing Letters, Vol. 29, March 2022, pp. 1247-1251.

[13] Zhao, Jing; Wang, Xiaomei; Li, Mingming. "Recent advances in self-supervised learning for audio processing." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 6, 2023, pp. 7182-7199.